

DOCUMENT RESUME

ED 270 471

TM 860 350

AUTHOR Cohen, Peter A.
TITLE An Updated and Expanded Meta-Analysis of Multisection Student Rating Validity Studies.
PUB DATE Apr 86
NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; Classroom Research; Class Size; College Faculty; Correlation; Course Evaluation; Effect Size; *Faculty Evaluation; Higher Education; Interrater Reliability; Literature Reviews; *Meta Analysis; Predictor Variables; Research Design; Sample Size; Self Evaluation (Individuals); Student Evaluation; *Student Evaluation of Teacher Performance; *Test Validity

ABSTRACT

A meta analysis was conducted of studies of college students' evaluations of their teachers. Forty-seven studies were included, representing 74 multisection courses, in which four selection criteria were met. An actual, rather than an experimental class was used; the unit of analysis was the class rather than the individual student; a common achievement measure was used across all class sections; and data were available to calculate the correlation between student achievement and ratings on instructors. The major outcomes of interest were the relationships between: (1) overall instructor ratings and student learning; (2) overall course ratings and student learning; (3) students' self-rating of learning and student achievement; and (4) student rating dimensions as identified commonly in the literature, and student learning. Twenty variables were used to categorize the research methodology. Achievement was indicated by final examination scores, cumulative point totals, or final grades. Results indicated a significant positive correlation between overall instructor rating and student learning. The size of this effect was different when teachers, rather than external graders, evaluated the exam; ratings were administered before students knew their final grade; or a small number of sections was studied. Both course ratings and student self-ratings were positively correlated with actual learning. References, categories used to describe study features, and extensive data tables are appended. (GDC)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED270471

An Updated and Expanded Meta-Analysis of
Multisection Student Rating Validity Studies

Peter A. Cohen, Ph.D.

Division of Instructional Development
The University of Texas
Health Science Center at San Antonio
7703 Floyd Curl Drive
San Antonio, Texas 78284

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- ☒ This document has been reproduced as
received from the person or organization
originating it.
- ☐ Minor changes have been made to improve
reproduction quality.

- Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

P. A. Cohen

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

Presented at the Annual Meeting of the
American Educational Research Association

San Francisco, 1986

TM 860 350

An Updated and Expanded Meta-Analysis of Multisection Student Rating Validity Studies¹

The general theme of this symposium is to provide "new" perspectives regarding the multisection validity literature, with a particular emphasis on recent developments. At least a dozen narrative reviews and three independent statistical, or meta-analytic reviews, have drawn conclusions about the relationship between student ratings and student learning. Why, then, are we focusing again on a literature that many would argue has been reviewed ad nauseam? Three reasons come to mind. First of all, the reviews and in particular the meta-analytic reviews have reached different conclusions regarding the validity of student ratings. Second, student ratings by nature are controversial. We've all heard the arguments posed by faculty: students are not qualified to make judgments on teaching competence, student ratings are nothing more than popularity contests, student ratings are unreliable and invalid, student ratings are highly related to grades students receive, extraneous factors and conditions affect student ratings, and so forth. And third, although there are inherent weaknesses in the generalizability of the design, multisection studies may be the best evidence we have to address the validity question. Commenting on the multisection validity approach Marsh (1984) states, "Perhaps more than any other type of study, the credibility of student ratings has rested on this paradigm."

The interest in relating student ratings to learning using the multisection validity approach is by no means a recent development. Remmers correlated student rating and achievement data almost 60 years ago (Remmers, 1928), and he published the first multisection validity study 37 years ago (Remmers, Martin, and Elliot, 1949). Some 30 years after this study, I attempted to locate and statistically synthesize results from all published and unpublished multisection validity studies that met certain specified criteria (Cohen, 1981). This morning, I will share with you the results of an updated and expanded meta-analysis of this literature that Philip Abrami and I have conducted recently.

Method

The approach we used in this meta-analysis was similar to that used in the earlier study (Cohen, 1981). We followed procedures described by Glass, McGaw, and Smith (1981) that require a reviewer to: a) locate studies in an area through objective and replicable procedures; b) code studies for relevant study features; c) code study outcomes using a common metric; and d) use statistical procedures to relate study outcomes and features.

Data Sources

For the original analysis, 41 studies reporting on 68 multisection courses were located. We performed extensive search procedures to update and supplement the original literature search. Ten separate computer-searches [Psychological Abstracts, ERIC, SSCI-Centra, SSCI-Cohen, SSCI-subject search (1972-1976), SSCI-subject search free text (1977-1985), SOCA-achievement, evaluation, and higher education (1963-1985), Medline-achievement, evaluation, and higher education (1971-1985), NTIS-achievement, evaluation, and higher education, and INSPEC (1970-1985)] yielded 60 potentially useful studies. By branching from bibliographies of located studies and manually searching recent issues of journals such as Educational and Psychological Measurement, Journal of Educational Psychology, and Research in Higher Education, we increased the number of newly located studies to 128.

To be included in the final sample, a study had to meet four general criteria. First, the study had to provide data from actual college or university classes. In other words, the study had to be a "field" study rather than an experimental analogue of teaching. Second, the unit of analysis in the study had to be the class (instructor) rather than the individual student. Third, the study needed to use a multisection format with a common achievement measure assessing learning for all sections. And fourth, the study had to provide data from which a ratings/achievement correlation could be calculated. Descriptive reports that did not provide enough data for the calculation of a correlational effect size were excluded.

Using these criteria, six studies (Hazelton, 1980; McKeachie, Lin, and Mendelson, 1978; Morgan and Vasch  1978; Murray, 1983; Orpen, 1980; Soper, 1973) were added to the original set of studies. The revised sample, therefore, included 47 studies reporting on 74 multisection courses.

Outcome Measures

The major outcomes of interest were: a) the relationship between overall instructor ratings and student learning; b) the relationship between overall course ratings and student learning; c) the relationship between students' self-ratings of learning and student achievement; and d) the relationship between student rating dimensions commonly identified in the literature and student learning. These rating dimensions are described in the earlier paper (Cohen, 1981). Student achievement was most commonly measured by a final examination. In some studies, however, cumulative point totals and final grades were used as an index of student learning. These indices were considered valid as long as they were derived from objective achievement criteria.

For statistical analysis, outcomes had to be expressed on a common scale of measurement. The basic measure of effect size here was Pearson's product-moment correlation. For each rating dimension, mean class rating was correlated with mean class achievement. Procedures outlined by Glass (1978) were used to convert various summary statistics into product-moment correlations. Before conducting statistical analyses, we converted all r 's to Fisher's Z -scores. After performing the analysis, we transformed the Z -scores into the more interpretable correlation coefficients. Weighted estimates of correlations were calculated using procedures suggested by Hedges and Olkin (1985).

Study Features

A total of 20 variables were used to describe methodologies, settings, and publication features of the studies. These study feature variables are described in Table 1.

INSERT TABLE 1 ABOUT HERE

Results

Because all but one of the studies reported a correlation between an overall instructor rating and student learning, we were able to conduct a complete statistical analysis for this relationship. The analysis covered both the overall correlational effect and the relationship between correlational effect sizes and study features. Only overall effects were calculated for other rating dimensions because of the limited number of studies in these areas. For all rating dimensions, mean rating/learning correlations are presented in Table 2.

INSERT TABLE 2 ABOUT HERE

Overall Instructor Rating

Correlations between an overall instructor rating and student learning were available for 73 of the 74 multisection courses located for this meta-analysis. For 64 of the 73 courses, overall instructor rating correlated positively with student learning; for nine courses the correlations were negative. For 33 courses the correlation coefficient reached statistical significance, and in 32 of these courses it was significantly positive.

The correlational effect size provides a more precise picture of the relationship between overall instructor rating and student learning. The average correlation in the 73 courses was .44 ($SE = .06$). Cohen (1977) operationally defined correlational effects of this magnitude as between medium and large. The 95 percent confidence interval on the true population correlation ranged from .34 to .53. To determine if it makes sense to talk about a single underlying overall instructor/student learning correlation, a test for homogeneity of correlations across studies was performed (Hedges and Olkin, 1985). The sample correlations were homogeneous at the 5 percent level, $Q = 92.04$, $df = 72$, $p > .05$.

Overall Instructor Rating and Study Features

Although the relationship between overall instructor rating and student learning was moderately large in the typical study, correlational effects varied in magnitude from study to study. The largest positive correlation reported was .97 (Morgan and Vasché, 1978); the largest negative correlation was $-.80$ (Bendig, 1953). Separate analyses were conducted to determine whether studies that reported large effect sizes differed systematically from those which produced small effects (Table 3). The 20 study feature variables together accounted for 62 percent of the variance in overall instructor rating/student learning correlations. Three study features were significantly related to correlational effect size. Average correlations differed in studies where: a) teachers, as opposed to external graders, evaluated the achievement examination; b) ratings were administered before, rather than after, students knew their final grade; and c) a small, rather than a moderate or large, number of sections was used.

[INSERT TABLE 3 ABOUT HERE]

Teacher bias in evaluating achievement. The average correlation in studies where students' own instructors graded the examinations was significantly lower than in studies where an external evaluator graded students' examinations and in studies where this information was unavailable, $F(2, 70) = 3.90$, $p < .05$. The average correlation in the 10 studies where teachers graded examinations for their students was .09 ($SE = .24$); it was .44 ($SE = .07$) for the 32 studies using an external grader; and it was .53 ($SE = .09$) in 31 studies that could not be coded for this study feature.

Timing of ratings. The average correlation in studies where ratings were administered after students knew their final grade was significantly higher than the average correlation in studies where ratings were administered before students knew their grades and in studies where this information was unavailable, $F = 6.26$ ($2, 70$), $p < .01$. The average correlation in the four studies where students rated instructors after they knew their final grades was .86 ($SE = .19$); it was .38 ($SE = .06$) in the 61

studies where students rated instructors before they knew their final grades; and it was .50 (SE = .19) in 8 studies that could not be coded for timing of ratings.

Number of sections. The average correlation in studies using a small multisection course (less than 10 sections) was significantly higher than the average correlation in studies using a medium (10-19 sections) or large (more than 19 sections) multisection course, $F = 4.50 (2, 70)$ $p < .05$. The average correlation in 26 studies with a small multisection course was .60 (SE = .13); it was .30 (SE = .10) in 21 studies with a medium multisection course; and it was .36 (SE = .04) in 26 studies with a large multisection course.

Overall Course Rating

Correlations between an overall course rating and student learning were available for only 23 of the 74 multisection courses. For 21 of the 23 courses, overall course rating was positively correlated with student learning; the correlation was negative for two courses. A total of 12 of the 23 correlations were statistically significant and all 12 showed significant positive correlations. The average correlation for the 23 courses was .46 (SE = .11).

Self-Rating of Learning

Correlations between mean section self-ratings of learning and actual learning were found for 17 multisection courses. The correlation between self-ratings of learning and actual learning was positive for 11 courses; it was negative for four courses; and for two courses the self-ratings of learning/actual learning correlation was zero. The correlations were statistically significant for five courses, and in each case the correlation was significant in a positive direction. The average correlation between self-ratings of learning and actual learning was .47 (SE = .11).

Specific Rating Dimensions

Average rating/learning correlations were large for the dimensions of Skill, mean $r = .53$ (SE = .09); Structure, mean $r = .52$ (SE = .11); and Interaction, mean $r = .42$ (SE = .18). The average correlations were more modest for the dimensions of Rapport, mean $r = .31$ (SE = .06); Feedback, mean $r = .29$ (SE = .03); and Evaluation, mean $r = .23$ (SE = .10). For the Difficulty dimension, there was no relationship between student ratings and learning, mean $r = .01$ (SE = .07).

Weighted Estimates of Correlational Effect Sizes

For each dimension, weighted estimates of the population correlations were calculated (Hedges and Olkin, 1985). This seemed particularly appropriate because of the effect of section size on overall instructor rating/student learning correlations. The weighted estimates are presented in Table 4. For almost all the dimensions, weighted mean correlations were smaller than the correspondent unweighted means.

INSERT TABLE 4 ABOUT HERE

Discussion

Narrative reviews, and even some meta-analyses, can be subject to bias based on which studies in an area are included in the review. Such selective reviewing can be problematic in controversial research areas, such as is the case with the student rating literature. A carefully conducted meta-analysis, on the other hand, can be considered a scientific review of the literature. For the present updated meta-analysis on multisection validity studies, we attempted to document precisely the procedures and methods used for locating, coding, and analyzing studies.

Our extensive search strategies resulted in only six additional studies being added to the original data set. The obtained mean correlations, therefore, were quite similar to those reported in the earlier analysis. For the present study, however, we also computed weighted estimates of correlations. These estimates were generally smaller than the unweighted correlational effect sizes and their confidence intervals were smaller. These findings suggest that rating/learning correlations are moderate in size and fairly consistent across studies. This is especially true concerning the overall instructor and overall course ratings. We feel that these results support the validity of student ratings.

We were particularly interested in the relationship between study features and effect size. The 20 study feature variables accounted for over 60 percent of the variance in overall instructor rating/learning correlations. As a set, methodological study features explained the greatest amount of variance (53 percent), with setting study features and publication features explaining only a small amount of effect size variability (nine percent and two percent, respectively).

Three methodological study features were significantly

related to effect size. In the few studies where ratings were administered after students knew their final grades, the ratings/learning correlations were artificially high. Of course, in this situation the interpretation of this relationship becomes confounded. Are good teachers receiving high ratings or are students rewarding teachers who give high grades? When addressing the validity of student ratings, we should only consider the results from multisection studies where timing of ratings is controlled. For these studies, the average rating/learning correlation is .38, a medium-sized effect.

The correlation between overall instructor ratings and learning was much smaller in studies where students' own instructors, as opposed to an external evaluator, graded achievement examinations. The possibility of inconsistencies in grading practices among section instructors could have attenuated the rating/learning correlation. Again, we need to take more seriously studies exercising control for this potential methodological bias.

Section size affected the size of the correlational effect. Correlations were higher in studies with small numbers of sections. Many reviewers have concluded that results from studies using small numbers of sections are quite variable and perhaps difficult to interpret. Our results support that view. For studies using 20 or more sections, the average correlation was .36. This seems to be a more accurate reflection of the true effect.

There are a number of issues still to be explored. We need to reexamine inclusion criteria. Should we accept in the pool every multisection study and view our results as representative of studies in the area, or should we be more rigorous in formulating criteria so that our results more likely represent a "true" effect? This is a hotly debated aspect of meta-analysis in general, and particularly in this multisection validity area.

Finally, we want to perform a more extensive analysis of the relationships between study feature variables and effect size. A number of predictor variables hypothesized to affect ratings, learning, or the relationship between ratings and learning were excluded from the present analysis. These include such student characteristics as gender, personality, attitudes, and reasons for section selection; instructor characteristics such as personality and attitudes, and variables under an instructor's control; setting characteristics such as purpose of ratings, rating anonymity, class size; and methodological features such as correspondence between number of instructors and number of sections and differential incentive (between sections) for the final examination. Although it is difficult to code these study features, attempts at coding may provide useful information about the state of our knowledge.

References

- Bendig, A.W. (1953). Student achievement in introductory psychology and student ratings of the competence and empathy of their instructors. Journal of Psychology, 36, 427-433.
- Cohen, J. (1977). Statistical power analysis for the behavioral sciences (Revised edition). New York: Academic Press.
- Cohen, P.A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. Review of Educational Research, 51, 281-309.
- Glass, G.V. (1978). Integrating findings: The meta-analysis of research. In L. S. Shulman (Ed.), Review of research in education (Vol. 5). Itasca, IL: F. E. Peacock.
- Glass, G.V., McGaw, B., and Smith, M.L. (1981). Meta-analysis in social research. Beverly Hills, CA: Sage.
- Hazelton, A.E. (1980). A study of the validity of student ratings of college teaching assessed on a criterion of student achievement in a first course in calculus. Unpublished doctoral dissertation, The University of Florida.
- Hedges, L.V., and Olkin, I. (1985). Statistical methods for meta-analysis. Orlando, FL: Academic Press.
- Marsh, H.W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. Journal of Educational Psychology, 76, 707-754.
- McKeachie, W.J., Lin, Y., and Mendelson, C. (1978). A small study assessing teacher effectiveness: Does learning last? Contemporary Educational Psychology, 3, 352-357.
- Morgan, W., and Vaschē, J. (1978). An educational production function approach to teaching effectiveness and evaluation. Journal of Economic Education, Spring, 123-126.
- Murray, H. (1983). Low inference classroom teaching behaviors in relation to six measures of college teaching effectiveness. Proceedings of the conference on "The Evaluation and Improvement of University Teaching: The Canadian Experience," Montebello, Quebec.
- Orpen, C. (1980). Student evaluation of lecturers as an indicator of instructional quality: A validity study. Journal of Educational Research, 74, 5-7.

Remmers, H.H. (1978). The relationship between students' marks and student attitudes toward instructors. School and Society, 28, 759-760.

Remmers, H.H., Martin, F.D., and Elliot, D.N. (1949). Are student ratings of their instructors related to their grades? Purdue University Studies in Higher Education, 44, 17-26.

Soper, J.C. (1973). Soft research on a hard subject: Student evaluations reconsidered. Journal of Economic Education, Fall, 22-26.

Notes

1. This study could not have been undertaken without the efforts of Dr. Philip Abrami and his research assistants, Sylvia D'Apollonia, Gretchen Lowerison, and Helen Kandarakis of Concordia University.
2. The original studies are listed in Cohen (1981).

Table 1
Categories Used to Describe Study Features

Methodology

Subject Assignment

Random--Students were assigned randomly to sections.

Evidence of Equivalence--If nonrandom assignment, was there evidence that sections were equivalent in terms of students' initial ability.

No Evidence--No evidence was presented in report.

Control for Scoring Bias

Objective--Objective tests were used to measure student achievement.

Nonobjective--Subjective decisions had to be made in scoring tests.

Control for Test-Author Bias

Standardized--A standardized test was used as the criterion measure for student learning.

Local--A locally developed test was used as the criterion measure.

Control for Teacher Bias in Evaluating Achievement

External--An external evaluator graded achievement tests.

Teachers--Teachers graded achievement tests for their own class.

Control for Bias in the Rating Instrument

Standardized--A standardized rating instrument with an underlying factor structure was used to assess teaching effectiveness.

Nonstandardized--A single-item rating or a locally-developed form was used to assess teaching effectiveness.

Control for Initial Student Ability

Statistical Control--A part correlation was computed between

ratings and a residualized achievement score.

No Statistical Control--The rating/achievement validity coefficient was based on the correlation between student ratings and the raw achievement score.

Control for Prior Knowledge of the Instructor

No Knowledge--Students did not know which teachers were teaching different sections of a course prior to enrollment in a specific section.

Prior Knowledge--Students did know who was teaching which section.

Control for Timing of Ratings

Before Final Grade--Ratings were collected before students knew their final grade or examination score.

After Final Grade--Ratings were collected after students knew their final grade or examination score.

Duration of Instruction

A complete academic term.

Less than a complete term.

Teacher Autonomy

100%--Teachers were responsible for all classroom instruction.

Less than 100%--Teachers were responsible for only a component (e.g., discussion section) of instruction.

Number of Sections

The number of sections on which the rating/achievement correlations were based.

Study Quality

Based on other methodological variables, studies were classified as low, moderate, or high in study quality.

Settings

Course Emphasis on Science

"Hard" Science--Course emphasizes the hard sciences, engineering, mathematics, or agriculture.

"Soft" Discipline--Emphasis is on the social sciences, humanities, or education.

Course Emphasis on Pure Knowledge

Pure Knowledge--Course does not emphasize practical application of knowledge, e.g., English, chemistry, physiology, and psychology courses.

Applied Knowledge--Emphasis is on practical application, e.g., courses in mechanical engineering, special education, and economics.

Course Emphasis on Life Systems.

Life--Course emphasis is on living or organic objects of study, e.g., courses in physiology, special education, and psychology

Nonlife Systems--Emphasis is on inanimate objects, e.g., courses in chemistry, mechanical engineering, English, and economics.

Course Level

Introductory course

Advanced course

Institutional Setting

Doctorate-granting

Comprehensive, Liberal Arts, or Community College

Instructor Experience

Faculty--Course instructors were faculty members.

Faculty and Graduate Students--Faculty members and graduate students served as course instructors.

Graduate Students--Course instructors were graduate students.

Publication Features

Source of Study

Published--Professional journal article, scholarly book,
etc.

Dissertation

Unpublished--Technical report, clearinghouse document,
convention paper, etc.

Year of the Report

Before 1950

1950-1959

1960-1969

1970-1979

1980-1986

Table 2
Mean Rating/Learning Correlational Effect Sizes

Rating Dimension	N	Mean Correlation	95% Confidence Interval
Overall Instructor	73	.44	.34, .53
Overall Course	23	.46	.28, .62
Self-Rating of Learning	17	.47	.28, .63
Specific Rating Dimensions			
Skill	43	.53	.38, .65
Rapport	31	.31	.20, .42
Structure	30	.52	.34, .67
Difficulty	26	.01	-.13, .15
Interaction	16	.42	.05, .68
Feedback	7	.29	.21, .36
Evaluation	26	.23	.04, .41

Table 3

Means and Standard Errors of Overall Instructor Rating/Student Learning
Correlational Effect Sizes for 73 Multisection Courses
Classified by Study Features

Categories	N	Mean Correlation	Standard Error
1. Subject Assignment			
Random	15	.40	.07
Evidence of Equivalence	8	.42	.15
No Evidence	50	.45	.08
2. Scoring Bias			
Objective	43	.42	.08
Nonobjective	8	.28	.17
No Data	22	.53	.11
3. Test-Author Bias			
Standardized	7	.44	.18
Local	61	.41	.06
No Data	5	.72	.40
4. Teacher Bias*			
External	32	.44	.07
Teachers	10	.09	.24
No Data	31	.53	.09
5. Rating Instrument Bias			
Standardized	48	.46	.07
Nonstandardized	24	.33	.10
No Data	1	.97	0

Categories	N	Mean Correlation	Standard Error
6. Statistical Control for Ability			
Yes	26	.50	.13
No	47	.41	.06
7. Prior Knowledge of Instructor			
No Knowledge	28	.47	.08
Prior Knowledge	6	.39	.31
No Data	39	.42	.09
8. Timing of Ratings**			
Before Final Grades	61	.38	.06
After Final Grades	4	.86	.19
No Data	8	.50	.19
9. Duration of Instruction			
Complete Academic Term	71	.42	.06
Fraction of Academic Term	2	.72	.45
10. Teacher Autonomy			
100%	58	.45	.06
Less Than 100%	15	.40	.16
11. Number of Sections*			
More Than 19	26	.36	.04
10-19	21	.30	.10
Less Than 10	26	.60	.13

Categories	N	Mean Correlation	Standard Error
12. Study Quality			
High	7	.43	.05
Moderate	59	.40	.06
Low	7	.69	.38
13. Course Emphasis on Science			
Hard	37	.42	.07
Soft	36	.45	.10
14. Course Emphasis on Pure Knowledge			
Pure	65	.43	.07
Applied	8	.47	.24
15. Course Emphasis on Life Systems			
Life	30	.42	.08
Nonlife	43	.46	.08
16. Course Level			
Introductory	70	.42	.06
Advanced	3	.64	.20
17. Institutional Setting			
Doctorate-granting	49	.42	.08
Comprehensive, Liberal Arts, or Community College	24	.46	.07

Categories	N	Mean Correlation	Standard Error
18. Instructor Experience			
Faculty	41	.52	.07
Faculty and Graduate Students	5	.17	.40
Graduate Students	26	.37	.10
No Data	1	.14	0
19. Source of Study			
Published	58	.44	.07
Dissertation	1	.10	0
Unpublished	14	.45	.14
20. Year of Report			
Before 1950	1	.49	0
1950-1959	6	.25	.32
1960-1969	3	.75	.26
1970-1979	59	.44	.06
1980-1986	4	.34	.26

*Difference in category means significant for this study feature, $p < .05$.

**Difference in category means significant for this study feature, $p < .01$.

Table 4
Weighted Correlational Effect Sizes

Rating Dimension	N	Weighted Mean	95% Confidence Interval
Overall Instructor	73	.35	.30, .40
Overall Course	23	.37	.29, .45
Self-Rating of Learning	17	.34	.22, .46
Specific Rating Dimensions			
Skill	43	.42	.36, .48
Rapport	31	.27	.19, .35
Structure	30	.30	.20, .40
Difficulty	26	.03	-.08, .14
Interaction	16	.27	.16, .38
Feedback	7	.30	.16, .44
Evaluation	26	.24	.15, .33